CLAIMS

I claim:

1. A method for the automatic configuration of dynamic database search forms comprising:

obtaining a database listing containing the uniform resource locators (URLs) for each one of a plurality of databases to be configured;

accessing each one of said plurality of databases;

capturing a web page from the database associated with said URL;

locating data entry windows in said captured web page;

selecting a most probable data entry window of data entry windows for passing queries to said database;

searching candidate responses for a next link indicating a next page for additional results from said database in response to a query; and

writing an engine file describing the form layout and requirements based upon said candidate responses and said next link.

2. The method of claim 1, wherein the step of accessing each one of said plurality of databases further comprises accessing a network and following a URL to a database on said network to be configured for automatic completion of search forms.

3. The method of claim 1, wherein the step of locating data entry windows in said captured web page further comprises:

saving information captured from the web page as a source

version of the web page;

filtering said source version into additional listings of URLs and text portions;

examining said text portions for occurrences of a form label;

collecting each form tagged with the form label;

scoring each one of said forms to develop a numerical representation of a likelihood that any one form is a query input form;

selecting one of said forms based on said form having a higher numerical representation than any other of said forms;

storing an action string associated with said form, said action string comprising a URL having at least a domain portion, a program portion, and a query portion;

storing a form method indicator associated with said database.

4. The method of claim 3, wherein the step of scoring each one of said forms further comprises:

locating an action string associated with said data entry window;

obtaining a listing of bad action string;

comparing said action string with said listing of bad action strings and determining if a portion of said action string matches any bad action strings of said listing of bad action strings, setting said numerical representation to zero and terminating said step of scoring if a portion of said action string matches any of said bad action strings within a predefined window determined by a binding factor;

setting a name matching metric;

setting an undesirable link text metric;

setting an undesirable value metric;

setting a desirable link text metric;

setting a null text metric;

computing a said numerical representation.

5. The method of claim 4, wherein the step of setting said numerical representation further comprises using value of 0 for said binding factor associated with said bad action string metric.

6. The method of claim 4, wherein the step of setting a name matching metric further comprises:

locating said data entry URL associated with the data entry window;

locating a page URL associated with the web page;

comparing a host name portion of said data entry URL with a host name portion of said page URL;

setting a name matching metric to a presence predetermined value if said host name portion of said data entry URL matches said host name portion of said page URL;

setting a name matching metric to an absence predetermined value if said host name portion of said data entry URL does not match said host name portion of said page URL;

7. The method of claim 6, wherein said steps of setting a name matching metric further comprise:

using a value of 6 for said presence predetermined value associated with said name matching metric;

using a value of 0 for said absence predetermined value associated with said name matching metric.

8. The method of claim 4, wherein the step of setting an undesirable link text metric further comprises:

locating said action string associated with said data entry window;

obtaining a listing of undesirable link texts;

comparing said action string with said listing of undesirable link text and determining if a portion of said action string matches any undesirable link texts of said listing of undesirable link texts, setting said numerical representation to zero and terminating said step of scoring if a portion of said action string matches any of said undesirable link texts within a predefined window determined by a binding factor.

9. The method of claim 8, wherein said steps of setting an undesirable link metric further comprises using a value of 1 for said binding factor associated with said undesirable link text.

10. The method of claim 5, wherein the step of setting an undesirable value metric further comprises:

locating said action string associated with said data entry window;

obtaining a listing of undesirable values;

comparing said action string with said listing of undesirable value and determining if a portion of said action string matches any undesirable values of said listing of undesirable values, setting an undesirable value metric to a presence predetermined value if a portion of said action string matches any of said undesirable values within a predefined window determined by a binding factor, and setting an undesirable value metric to an absence predetermined value if a portion of said action string does not match an undesirable value within a predefined window determined by a binding factor.

11. The method of claim 10, further comprising:

using a value of 0 for said presence predetermined value

associated with said undesirable value metric;

using a value of 4 for said absence predetermined value associated with said undesirable value metric;

using a value of 0 for said binding factor associated with said undesirable value metric.

12. The method of claim 5, wherein the step of setting a desirable link text metric further comprises:

locating said action string associated with said data entry window;

obtaining a listing of desirable link texts;

comparing said action string with said listing of desirable link text and determining if a portion of said action string matches any desirable link texts of said listing of desirable link texts, setting an desirable link text metric to a presence predetermined value if a portion of said action string matches any of said desirable link texts within a predefined window determined by a binding factor, and setting an desirable link text metric to an absence predetermined value if a portion of said action string does not match an desirable link text within a predefined window determined by a binding factor.

13. The method of claim 12, further comprising:

using a value of 4 for said presence predetermined value associated with said desirable text metric;

using a value of 0 for said absence predetermined value associated with said desirable text metric;

using a value of 2 for said binding factor associated with said desirable text metric.

14. The method of claim 4, wherein the step of setting a null

text metric further comprises:

locating said action string associated with said data entry window;

checking said action string for an absence of associated text;

setting a null text metric to a presence predetermined value if no text is associated with said form.

15. The method of claim 14, wherein said step of setting a null text metric further comprises using a value of 2 for said null text metric.

16. The method of claim 4, wherein said step of calculating said numerical representation further comprises adjusting said numerical representation by adding an integer value determined by the number of edit boxes on said web page.

17. The method of claim 3, wherein the step of scoring each one of said forms further comprises:

locating an action string associated with said data entry window;

obtaining a listing of bad action strings;

comparing said action string with said listing of bad action strings and determining if a portion of said action string matches any bad action string of said listing of bad action strings, setting said numerical representation to zero and terminating said step of scoring if said bad action string matches a portion of said action string within a predefined window determined by a binding factor;

wherein the step of setting said numerical representation further comprises using value of 0 for said binding factor associated with said bad action string metric;

locating a page URL associated with the web page;

comparing a host name portion of said data entry URL with a host name portion of said page URL;

setting a name matching metric to a presence predetermined value if said host name portion of said data entry URL matches said host name portion of said page URL;

setting a name matching metric to an absence predetermined value if said host name portion of said data entry URL does not match said host name portion of said page URL;

said steps of setting a name matching metric further comprises:

using a value of 6 for said presence value associated with said name matching metric;

using a value of 0 for said absence value associated with said name matching metric;

obtaining a listing of undesirable link texts;

comparing said action string with said listing of undesirable link text and determining if a portion of said action string matches any undesirable link texts of said listing of undesirable link texts, setting said numerical representation to zero and terminating said step of scoring if a portion of said action string matches any of said undesirable link texts within a predefined window determined by a binding factor;

using a value of 1 for said binding factor associated with said undesirable link text;

obtaining a listing of undesirable values;

comparing said action string with said listing of undesirable value and determining if a portion of said action string matches any undesirable values of said listing of undesirable values, setting an undesirable value metric to a presence predetermined value if a portion of said action string matches any of said undesirable values

within a predefined window determined by a binding factor, and setting an undesirable value metric to an absence predetermined value if a portion of said action string does not match an undesirable value within a predefined window determined by a binding factor;

using a value of 0 for said presence predetermined value associated with said undesirable value metric;

using a value of 4 for said absence predetermined value associated with said undesirable value metric;

using a value of 0 for said binding factor associated with said undesirable value metric;

obtaining a listing of desirable link texts;

comparing said action string with said listing of desirable link text and determining if a portion of said action string matches any desirable link texts of said listing of desirable link texts, setting an desirable link text metric to a presence predetermined value if a portion of said action string matches any of said desirable link texts within a predefined window determined by a binding factor, and setting an desirable link text metric to an absence predetermined value if a portion of said action string does not match an desirable link text within a predefined window determined by a binding factor;

using a value of 4 for said presence predetermined value associated with said desirable text metric;

using a value of 0 for said absence predetermined value associated with said desirable text metric;

using a value of 2 for said binding factor associated with said desirable text metric;

checking said action string for an absence of associated text;

setting a null text metric to a presence predetermined value if

no text is associated with said form;

using a value of 2 for said null text metric;

computing a numerical representation of the likelihood that said data entry is a correct data entry window for passing queries to said database; and

adjusting said numerical representation by adding an integer value determined by the number of edit boxes on said web page.

18. The method of claim 1, further comprising:

determining a location of each one of a plurality of results locations on a responsive web page where results from a query are posted;

determining a location of each one of a plurality of non-results items on a responsive page are posted;

19. The method of claim 18, further comprising:

selecting a plurality of validation queries;

submitting a first one of said plurality of validation queries to said database using said action string;

capturing a first responsive web page returned in response to said first one of said plurality of validation queries;

resubmitting said first one of said plurality of validation queries to said database using said action string;

capturing a second responsive web page returned in response to said second submission of said first one of said plurality of validation queries;

comparing said first responsive web page with said second responsive web page, any differences between said first and second responsive web page are extraneous responses and are ignored, storing any duplicates between said first and second responsive web pages as candidate responses to said validation query;

storing said candidate responses;

submitting a second one of said plurality of validation queries to said database using said action string;

capturing a responsive web page returned in response to said second validation query;

repeating submission of additional validation queries and capture of additional responsive web pages until all validation queries have been submitted;

comparing said first responsive web page to each of said additional responsive web pages, ignoring any duplicates between said first responsive and additional responsive web pages as extraneous responses, storing any differences between said first responsive and said additional responsive web pages as candidate responses to said validation query;

comparing each one of said responsive web pages to all other said responsive web pages, ignoring any duplicates between said responsive web pages as extraneous responses, storing any differences between said responsive web pages as candidate responses to said validation query; and

searching candidate responses for a next link indicating a next page for additional results from said database in response to said query.

20. The method of claim 19, wherein said step of comparing said first responsive web page with said second responsive web page further comprises:

comparing each URL in said first responsive web page with each URL in said second responsive web page;

capturing a location associated with every URL common between said first responsive web page and said second responsive web page as a potential location for results from a query;

capturing a location associated with every URL distinct between said first responsive web page and said second responsive web page as a potential location not associated with results from a query;

comparing each label associated with each URL in said first responsive web page with each label associated with each URL in said second responsive web page;

capturing a location, associated with every label associated with every URL, which is common between said first responsive web page and said second responsive web page as a potential location for results from a query;

capturing a location, associated with every label associated with every URL, which is distinct between said first responsive web page and said second responsive web page as a potential location not associated with results from a query.

21. The method of claim 19, wherein said step of comparing said first responsive web page with said additional responsive web page further comprises:

comparing each URL in said first responsive web page with each URL in said additional responsive web page;

capturing a location associated with every URL common between said first responsive web page and said additional responsive web page as a potential location for results from a query;

capturing a location associated with every URL distinct between said first responsive web page and said additional responsive web page as a potential location not associated with results from a query;

comparing each label associated with each URL in said first responsive web page with each label associated with each URL in

said additional responsive web page;

capturing a location, associated with every label associated with every URL, which is common between said first responsive web page and said additional responsive web page as a potential location for results from a query;

capturing a location, associated with every label associated with every URL, which is distinct between said first responsive web page and said additional responsive web page as a potential location not associated with results from a query.

22. The method of claim 19, wherein said step of comparing each one of said responsive web pages with all other said responsive web pages further comprises:

comparing each URL in each one of said responsive web pages with each URL in all other said responsive web pages;

capturing a location associated with every URL common between each one of said responsive web pages and all other said responsive web pages as a potential location for results from a query;

capturing a location associated with every URL distinct between each one of said responsive web pages and all other said responsive web pages as a potential location not associated with results from a query;

comparing each label associated with each URL in each one of said responsive web pages with each label associated with each URL in all other said responsive web pages;

capturing a location, associated with every label associated with every URL, which is common between each one of said responsive web pages and all other said responsive web pages as a potential location for results from a query;

capturing a location, associated with every label associated

with every URL, which is distinct between each one of said responsive web pages and all other said responsive web pages as a potential location not associated with results from a query.

23. The method of claim 19, further comprising:

comparing each URL in said first responsive web page with each URL in said second responsive web page;

capturing a location associated with every URL common between said first responsive web page and said second responsive web page as a potential location for results from a query;

capturing a location associated with every URL distinct between said first responsive web page and said second responsive web page as a potential location not associated with results from a query;

comparing each label associated with each URL in said first responsive web page with each label associated with each URL in said second responsive web page;

capturing a location, associated with every label associated with every URL, which is common between said first responsive web page and said second responsive web page as a potential location for results from a query;

capturing a location, associated with every label associated with every URL, which is distinct between said first responsive web page and said second responsive web page as a potential location not associated with results from a query;

comparing each URL in said first responsive web page with each URL in said additional responsive web page;

capturing a location associated with every URL common between said first responsive web page and said additional responsive web page as a potential location for results from a query;

capturing a location associated with every URL distinct between said first responsive web page and said additional responsive web page as a potential location not associated with results from a query;

comparing each label associated with each URL in said first responsive web page with each label associated with each URL in said additional responsive web page;

capturing a location, associated with every label associated with every URL, which is common between said first responsive web page and said additional responsive web page as a potential location for results from a query;

capturing a location, associated with every label associated with every URL, which is distinct between said first responsive web page and said additional responsive web page as a potential location not associated with results from a query;

comparing each URL in each one of said responsive web pages with each URL in all other said responsive web pages;

capturing a location associated with every URL common between each one of said responsive web pages and all other said responsive web pages as a potential location for results from a query;

capturing a location associated with every URL distinct between each one of said responsive web pages and all other said responsive web pages as a potential location not associated with results from a query;

comparing each label associated with each URL in each one of said responsive web pages with each label associated with each URL in all other said responsive web pages;

capturing a location, associated with every label associated with every URL, which is common between each one of said

responsive web pages and all other said responsive web pages as a potential location for results from a query;

capturing a location, associated with every label associated with every URL, which is distinct between each one of said responsive web pages and all other said responsive web pages as a potential location not associated with results from a query.

24. The method of claim 19, wherein said step of selecting a plurality of validation queries further comprises:

selecting the term "home" as a first one of said plurality of validation queries;

selecting the term "copyright" as a second one of said plurality of validation queries;

selecting the term "web" as a third one of said plurality of validation queries.

25. The method of claim 19, wherein said step of selecting a plurality of validation queries comprises choosing three terms common to a subject matter area, with minimal overlap between at least two of the terms, for said validation queries.

26. The method of claim 19, wherein said step of searching candidate responses for a next link further comprises:

obtaining a next term listing providing a plurality of labels commonly associated with data entry windows used for accessing additional results from a database associated with a user's query;

comparing each label associated with each URL in said first

responsive web page with each one of said plurality of labels in order provided in said next term listing;

selecting a data entry window as a next link if said label associated with said data entry window matches one of said plurality of labels provided by said next link listing within a predetermined window defined by a binding factor.

27. The method of claim 26, wherein said step of selecting a data entry window further comprises:

determining if a match has been made between said label associated with said data entry window and one of said plurality of labels provided by said next link listing;

comparing each label associated with each URL in a first one of said additional responsive web pages associated with a second one of said validation queries with each one of said plurality of labels in order provided in said next term listing in no match has been made;

selecting a data entry window as a next link if said label associated with said data entry window matches one of said plurality of labels provided by said next link listing within a predetermined window defined by a binding factor if no prior match has been made.

28. The method of claim 26, wherein said step of selecting a data entry window further comprises using a value of approximately 1.5 for said binding factor.

29. The method of claim 3, wherein the step of scoring each one of said forms further comprises:

locating an action string associated with said data entry window;

obtaining a listing of bad action string;

comparing said action string with said listing of bad action strings and determining if a bad action string matches a portion of said action string; setting said numerical representation to zero and terminating said step of scoring if said bad action string matches a portion of said action string within a predefined window determined by a binding factor;

using a value of 0 for said binding factor associated with said bad action string metric;

locating said data entry URL associated with the data entry window;

locating a page URL associated with the web page;

comparing a host name portion of said data entry URL with a host name portion of said page URL;

setting a name matching metric to a presence predetermined value if said host name portion of said data entry URL matches said host name portion of said page URL;

setting a name matching metric to an absence predetermined value if said host name portion of said data entry URL does not match said host name portion of said page URL;

using a value of 0 for said absence predetermined value associated with said name matching metric;

obtaining a listing of undesirable link texts;

comparing said action string with said listing of undesirable link text and determining if a portion of said action string matches any undesirable link texts of said listing of undesirable link texts, setting said numerical representation to zero and terminating said step of scoring if a portion of said action string matches any of said undesirable link texts within a predefined window determined by a binding factor;

using a value of 1 for said binding factor associated with said undesirable link text;

obtaining a listing of undesirable values;

comparing said action string with said listing of undesirable value and determining if a portion of said action string matches any undesirable values of said listing of undesirable values, setting an undesirable value metric to a presence predetermined value if a portion of said action string matches any of said undesirable values within a predefined window determined by a binding factor, and setting an undesirable value metric to an absence predetermined value if a portion of said action string does not match an undesirable value within a predefined window determined by a binding factor;

using a value of 0 for said presence value associated with said undesirable value metric;

using a value of 0 for said binding factor associated with said undesirable value metric;

obtaining a listing of desirable link texts;

comparing said action string with said listing of desirable link text and determining if a portion of said action string matches any desirable link texts of said listing of desirable link texts, setting an desirable link text metric to a presence predetermined value if a portion of said action string matches any of said desirable link texts within a predefined window determined by a binding factor, and setting an desirable link text metric to an absence predetermined value if a portion of said action string does not match an desirable link text within a predefined window determined by a binding factor;

using a value of 0 for said absence predetermined value associated with said desirable text metric;

using a value of 2 for said binding factor associated with said desirable text metric;

checking said action string for an absence of associated text;

setting a null text metric to a presence predetermined value if no text is associated with said form;

using values for said presence predetermined value associated with said name matching metric, said absence predetermined value associated with said undesirable value metric, said presence predetermined value associated with said desirable text metric, and said null text metric such that the relative weighting of each of said metrics is approximately 3:2:2:1 respectively; and

computing a said numerical representation.

30. A system for the automatic configuration of dynamic database search forms comprising:

a computer system having a storage means for facilitating the retention and recall of dynamic database content, said computer system having a communications means for performing bi-directional communications between said computer system and a network;

a query input means for receiving a plurality of queries from a user and transferring the plurality of queries to a plurality of databases;

an action string module interfaced to said computer system for determining a format associated with an entry page for a database, said action string module being for determining an appropriate data entry window for use in passing a query to said database;

a results module interfaced to said computer system and said action string module, said results module locating areas on a responsive page returned by said database in response to said query

where results are placed;

a next link module interface to each one of said computer system, action string module, and results module, said next link module locating a link associated with additional results provided by said database in response to said query;

an engine file module interfaced to said computer system and every other module for storing results produced by each module such that a general format query is translatable into a database specific format allowing a common query to be submitted to multiple databases each requiring different formats.

31. The system of claim 30, further comprising a data comparison portion providing user specific information to each of said modules for facilitating analysis of said databases.

32. The system of claim 31, wherein said data comparison portion further comprises:

a database listing providing a URL for each of said databases to be analyzed;

a bad action string listing providing URLs for known databases which are not to be included in the analysis of said databases;

a desirable text link listing providing a plurality of desirable terms for use in analysis of said databases, a presence of any one of said plurality of desirable terms increasing a score associated with a data entry window on one of said responsive pages;

an undesirable text link listing providing a plurality of undesirable terms for use in analysis of said databases, a presence of any one of said plurality of undesirable terms setting a score associated with a data entry window on one of said responsive pages to 0 and ending analysis of said data entry window; and

an undesirable value listing providing a plurality of undesirable values for use in analysis of said databases, a presence of any one of said plurality of undesirable values decreases a score associated with a data entry window on one of said responsive pages.

33. The system of claim 31, wherein said data comparison portion further comprises:

a next link listing providing said next link module with a plurality of candidate terms for facilitating selection of a URL associated with a link to additional responses provided by said database in response to said query.

34. A system for the automatic configuration of dynamic database search forms comprising:

a computer system having a storage means for facilitating the retention and recall of dynamic database content, said computer system having a communications means for performing bi-directional communications between said computer system and a network;

a query input means for receiving a plurality of queries from a user and transferring the plurality of queries to a plurality of databases;

an action string module interfaced to said computer system for determining a format associated with an entry page for a database, said action string module being for determining an appropriate data entry window for use in passing a query to said database;

a results module interfaced to said computer system and said action string module, said results module locating areas on a responsive page returned by said database in response to said query

where results are placed;

a next link module interface to each one of said computer system, action string module, and results module, said next link module locating a link associated with additional results provided by said database in response to said query;

an engine file module interfaced to said computer system and every other module for storing results produced by each module such that a general format query is translatable into a database specific format allowing a common query to be submitted to multiple databases each requiring different formats;

a database listing providing a URL for each of said databases to be analyzed;

a bad action string listing providing URLs for known databases which are not to be included in the analysis of said databases;

a desirable text link listing providing a plurality of desirable terms for use in analysis of said databases, a presence of any one of said plurality of desirable terms increases a score associated with a data entry window on one of said responsive pages;

an undesirable text link listing providing a plurality of undesirable terms for use in analysis of said databases, a presence of any one of said plurality of undesirable terms sets a score associated with a data entry window on one of said responsive pages to 0 and ending analysis of said data entry window; and

an undesirable value listing providing a plurality of undesirable values for use in analysis of said databases, a presence of any one of said plurality of undesirable values decreases a score associated with a data entry window on one of said responsive pages;

a next link listing providing said next link module with a

plurality of candidate terms for facilitating selection of a URL associated with a link to additional responses provided by said database in response to said query.